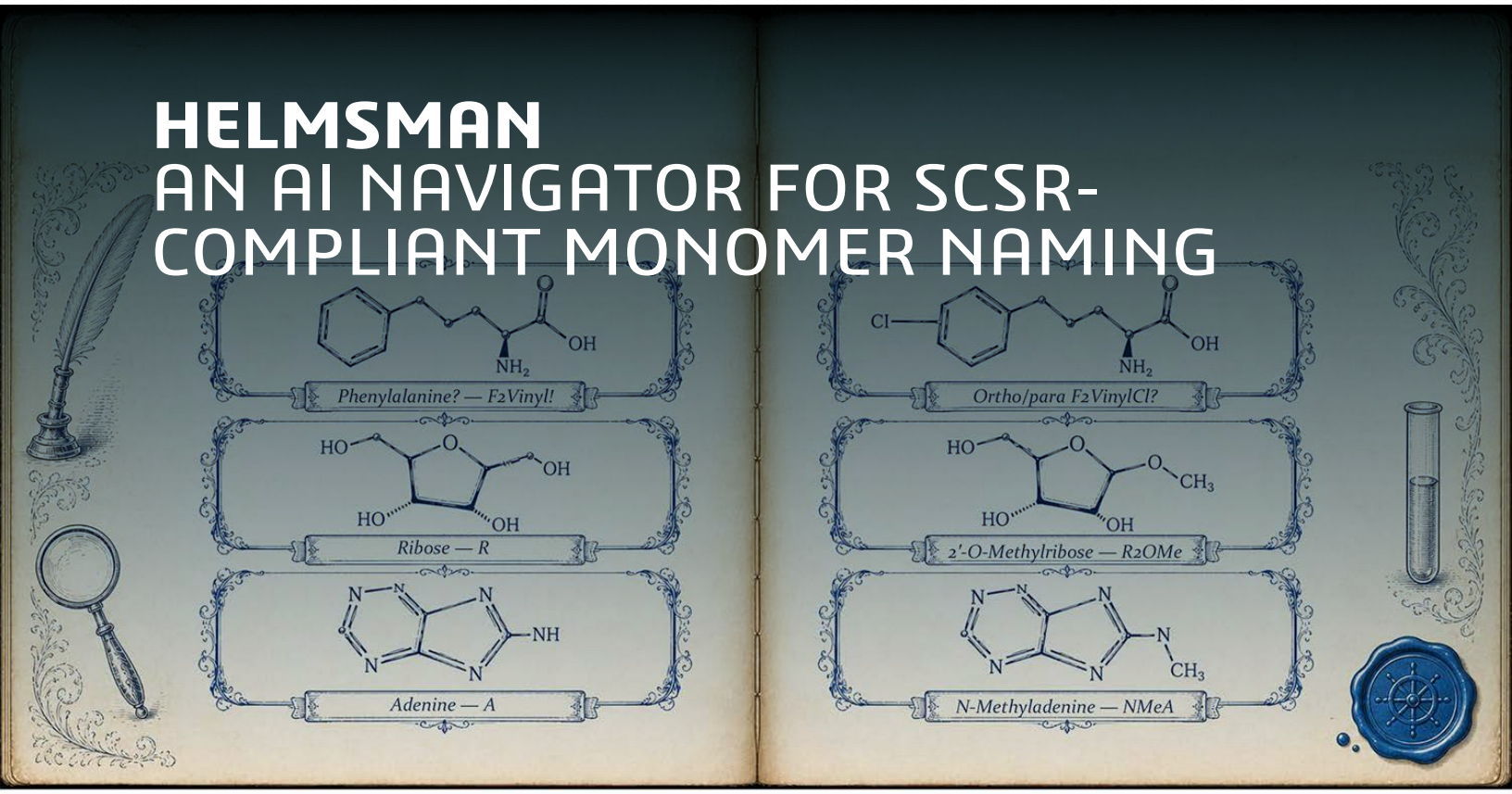
BIOVIA

3DEXPERIENCE®



HELMSMAN

an AI navigator for SCSR-compliant monomer naming

Arman Sadybekov · 2026

We present *Helmsman*, an automated pipeline for SCSR-compliant monomer naming that combines a deterministic rule engine with a three-role agentic ensemble (Curator, Reviewer, Scaffold Resolver) over a shared round-trip verifier and a final self-anchor backstop that registers irreducible residues as their own anchors. The rule engine adopts the naming conventions of HELMify (Godfrey et al., 2026); the ensemble extends the approach in two directions — out from amino acids to the full HELM monomer space, and from a single model-assisted pass to a structured escalation in which each role is given a progressively wider freedom to leave the rules' frame, bounded at every step by the same verifier. On a benchmark of 1,352 valid single-monomer inputs deliberately drawn from chemistry beyond today's curated HELM libraries (mined from a ≈ 2,800-entry hybrid mine, deduplicated against the bundled HELM Core Library and filtered to single-residue targets by a deterministic SMARTS pre-filter that removes sugar+base nucleosides, peptides, and other multi-monomer constructions), the rule engine alone names 478 (35.4 %); the rule engine plus the agentic loop names every valid input — **1,352 of 1,352** — a 100 % output-name guarantee underwritten by the self-anchor backstop and verified by structural read-back at every step.

DASSAULT SYSTEMES

THE CHALLENGE

Biopolymer chemistry has been outpacing its own vocabulary for sixty years. The IUPAC-IUB one-letter codes, the MODOMICS dictionary of modified nucleosides, HELM, and the SCSR molfile extension were each published behind the chemistry they had to encode (Chen et al., 2011; Dunin-Horkawicz et al., 2006; IUPAC-IUB Commission on Biochemical Nomenclature, 1968; Zhang et al., 2012); each closed the gap by adding a new layer of governed vocabulary on top of the previous one. The notations are now sufficient. The limit is the curation throughput of any manually maintained monomer catalogue, and the absence of an automated path from a chemical structure to a registration-grade name for a residue that is not yet catalogued.

The starting point for this work is **HELMify** (Godfrey et al., ChemRxiv 2026) (Godfrey et al., 2026), a recent and groundbreaking demonstration that a language model can be used to propose HELM-compatible names for monomers the curated catalogue does not cover. HELMify established the approach for amino acids; we asked how much further the approach could be taken. On a benchmark of 1,352 valid single-monomer inputs deliberately drawn from beyond the curated HELM libraries, Helmsman names every input — 1,352 of 1,352 — providing a **100 % naming guarantee** underwritten by a final self-anchor backstop that registers residues too singular to decompose against any existing parent as their own anchors.

WHY NAMING MATTERS

Two readings of the same problem make the case for an automated solution: one from inside Pistoia, showing how ad-hoc naming works and where it stops; one from outside, showing where a purely rule-based pipeline leaves gaps that a naïve LLM call cannot safely close.

Two residues from Pistoia

- **fl5C = 5-fluorocytosine:** Parses unambiguously: fl (fluoro) 5 (position 5) C (cytosine). Cytosine has well-defined atom positions; the substituent is a single halogen; nothing in the library competes for the name. Any compound that should be named fl5C is the same compound. Ad-hoc naming works because the answer is forced.
- **Tyr_Me = O-methyltyrosine:** Pistoia uses the symbol for the residue whose phenolic hydroxyl carries a methyl. Read as a HELM-style symbol, "Tyr with a methyl substituent" admits several chemically distinct interpretations — α-methyl-tyrosine, β-methyl-tyrosine, ring-methyl tyrosine, N-methyl tyrosine, and the catalogue's actual choice, O-methyl tyrosine. At least one alternative — α-methyl-tyrosine, or metyrosine, an FDA-approved tyrosine-hydroxylase inhibitor — is a real registration target whose chemistry is distinct from the catalogued structure and whose ad-hoc name would, by the same convention, also be Tyr_Me. The collision is silent. When the chemistry is forced, ad-hoc symbols are excellent; when it is not, they offer no protection against the naming collisions they invite.

The Chemist's Intuition—and the Pipeline it Leads To

The chemist's intuitive route from a structure to a name is the same each time: identify the closest familiar parent, name the modifications relative to that parent, concatenate the result. The parent supplies a stable origin; the modifications supply the differentiating descriptors; the combination is back-parseable when the parent set, the modification vocabulary, and the ordering conventions are all agreed in advance. HELMify (Godfrey et al., 2026) was the first published demonstration that this intuition can be automated end-to-end. Helmsman decomposes the intuition into three ingredients implemented and audited independently — **structural anchors** (canonical parents), **common-substructure decomposition** against the chosen parent, and a **substituent catalogue plus ordering rules** (the third adopted from HELMify unchanged). One consequence makes the rest of the paper tractable: every name the pipeline emits is built from a vocabulary the parser already knows, so the parser can read the name back, the reconstructed structure can be canonicalised, and the two canonicalisations can be compared. That round-trip verifier is the keystone of everything that follows.

Two residues that are not in Pistoia Alliance Core HELM Library

The two examples below were chosen because they expose the two failure modes of a purely rule-based version of the pipeline. Both are recently published residues with primary-literature pedigree; neither appears in the bundled HELM Core Library.

- **5-Carboxycytosine (5caC)** is the most oxidised TET-cycle intermediate in active DNA demethylation (He et al., 2011); the sibling 5hmC and 5fC bases are catalogued, 5caC is not. Helmsman picks cytosine by exact substructure match; decomposition reports a single additive modification — a carboxyl at C5 (Figure 1). The rules cannot finish because the catalogue has no substituent name for the carboxyl group. The chemistry is within the spirit of the pipeline; the catalogue is one entry short. A naïve LLM call asked to "name this residue" might supply CYT5COOH, CYT5CO2H, or CYT5carboxyl, with no guarantee that two calls produce the same answer or that the answer parses back into the original structure.

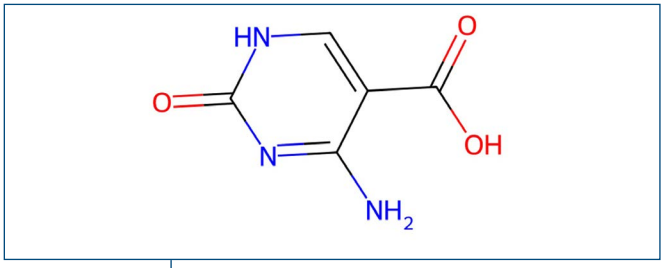


Figure 1: 5-Carboxycytosine (5caC), free base.

- **β-Hydroxyenduracididine (βhEnd)** is the signature residue of the mannopeptimycin antibiotics (Cochrane & Vederas, 2016; Wang et al., 2016). The anchor selector picks threonine, not arginine, because the Cα / β-hydroxyl / carboxyl patch is atom-for-atom more threonine-like (Figure 2); under threonine, the defining cyclic guanidine cannot be expressed by any combination of catalogue substituents. A naïve LLM call asked to complete the name under this wrong anchor will produce a string of the right shape — but the implied chemistry does not reconstruct βhEnd. The right move is not to invent a substituent name; it is to revisit the choice of parent scaffold.

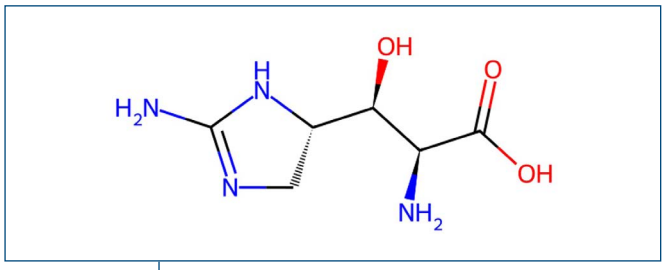


Figure 2: β-Hydroxyenduracididine (βhEnd), L-form free amino acid.

5caC is what the rule engine almost names — only a missing substituent stands between it and a verified output. βhEnd is what the rule engine misframes — no amount of added substituent chemistry would have fixed it. Closing both requires distinct interventions sharing the same baseline pipeline.

HELMSMAN: RULES, AN ENSEMBLE, AND A VERIFIER

Helmsman complements the Pistoia HELM Core Monomer Set rather than replacing it. Catalogued residues are returned under their catalogued symbols. Residues that are not catalogued are processed in two layers — first by the deterministic rule engine sketched above, and then, if any modification is left unresolved, by an agentic ensemble whose three roles share the same governed vocabulary and the same structural verifier.

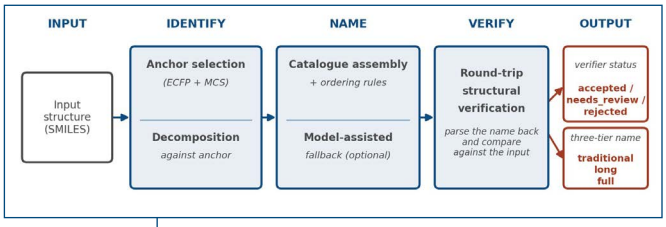


Figure 3: Helmsman pipeline overview.

The pipeline runs in three stages — identify, name, verify (Figure 3). The rule engine matches an input against the closest of 28 canonical anchor scaffolds, enumerates the structural differences, matches each one against a catalogue of expert-validated substituent tokens under the HELMify ordering conventions, and emits a name. The verifier parses the name back, canonicalises the implied structure, and compares it to the input. The verifier returns accepted (the structures match exactly), *needs_review* (the structural skeleton matches but the name uses a substituent or parent scaffold not yet in the curated catalogue, providing a controlled promotion path), or rejected (the parse fails or the structures disagree).

The Agentic Loop—Three Agents, the Rules, One Structural Check

When the rule engine cannot name every modification, the residues that remain are handled by an **agentic loop** in which three agents — **Curator** (also called the substituent Resolver), **Reviewer**, and **Scaffold Resolver** — work alongside the rules engine in repeating cycles until no further names can be added (Figure 4). This is the same set of agents an earlier single-pass version of the pipeline used; the change is that the same steps now repeat until naming is exhausted, rather than running once and stopping.

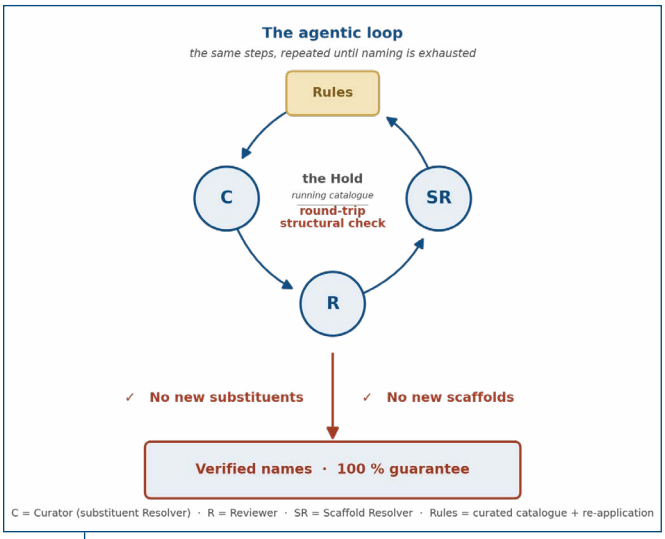


Figure 4: The agentic loop. The rules engine and three agents cycle over the Hold (a running catalogue of every named substituent, Reviewer verdict, and registered scaffold) and a round-trip structural check that every new entry must pass. The loop stops when two conditions hold simultaneously: there are no new substituents for the Curator to name AND a fresh Scaffold Resolver pass yields no new parent scaffolds. A small whole-molecule backstop afterwards closes the gap to a 100 % naming guarantee for every valid single-monomer input.

The default pass — name the missing substituents. A simple, deterministic check identifies the residues the rules have not yet named. The **Curator** proposes a name for the unnamed substituent on each one. Each proposed name is reviewed immediately by the Reviewer — kept, revised, or removed — and the rules then re-apply across the whole library with whatever substituent names survived the audit, so the gain propagates to every sibling residue carrying the same substituent. The pass repeats until no unnamed substituents remain. This is the name-once-reuse-everywhere property: the language model is consulted once per distinct piece of chemistry, not once per molecule.

The escalation pass — Scaffold Resolver, once. When no further substituent naming is possible, the **Scaffold Resolver** takes a single pass over whatever still remains. It may **re-match the residue** against a different parent scaffold from the extended scaffold library (98 vs the base 28); **add a new parent scaffold** to the library when the residue contains a recurring novel backbone (added only if that backbone is a substructure of the molecule that triggered it); or **assemble a name directly** under a relaxed structural check that requires the right backbone, the right modifications, and the right placement, with the name flagged for review afterwards. The rules re-apply again. If any of these moves yields newly named residues, the loop re-enters the default pass.

The stopping condition. The loop stops only when **both** conditions hold simultaneously: there are no new substituents for the Curator to name and a fresh Scaffold Resolver pass yields no new parent scaffolds. The dual condition guarantees the loop did not stop with easy naming left to do, and did not stop while an escalation step was still finding names.

The backstop. After the loop has stopped, a small **whole-molecule backstop** records any truly irreducible residue — a molecule that cannot be expressed as a parent scaffold plus substituents — as its own scaffold. The resulting one-off entry is verified by the same structural check that screens every other entry, and it closes the gap to a 100 % naming guarantee for every valid single-monomer input.

The shared catalogue and the structural check. All four participants in the loop share two things. The Hold is a persistent catalogue — every named substituent, every Reviewer verdict, every registered scaffold is recorded there — and the rules consult it on every pass, so a new entry by any agent is immediately available to all the others. The **round-trip structural check** is the universal screen that every entry must pass: the proposed name is parsed back into a molecular structure, and the entry is admitted only if that structure is identical to the input.

Three-Tier Name Output

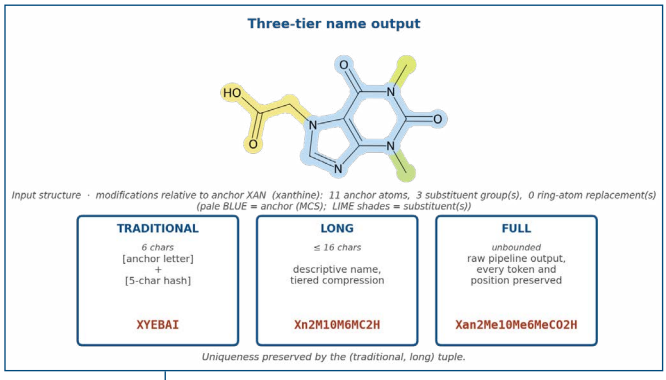


Figure 5: The three-tier name output.

Every accepted input is recorded with three name forms (Figure 5): a fixed **traditional** name (anchor first letter + 5-character deterministic hash) for column-bounded display; a descriptive **long** name with tiered compression (≤ 16 characters); and the **full** name with no length cap. Uniqueness across registry entries is preserved by the (traditional, long) tuple.

RESULTS

The pipeline was evaluated on a benchmark of 1,352 valid single-monomer residues, drawn from a ≈ 2,800-entry hybrid mine of MODOMICS (Boccaletto et al., 2022) (modified RNA residues), SwissSidechain (Gfeller et al., 2013) (non-natural amino-acid side chains), and substructure slices of ChEMBL (Zdrazil et al., 2024) (natural products, α-amino-acid backbones, nucleobase cores), de-duplicated against the bundled HELM Core Library. Roughly half of the mined candidates are not valid single-monomer inputs and are removed up front: a deterministic SMARTS-based detector identifies sugar+base nucleosides (~ 400 entries — HELM represents these as separate base + sugar [+ phosphate] residues, not single monomers), and complementary rules cut peptide bonds, di-residues, oversize entries, and unparsable structures (~ 950 further entries). The clean 1,352-entry single-monomer benchmark is what feeds the pipeline. Database portals were accessed on 2026-05-29.

The rule engine alone names **478 / 1,352 (35.4 %)**; the rule engine plus the agentic loop names **every input — 1,352 of 1,352.** The operative claim is a structural one: the pipeline returns a verified, registration-grade name for every valid single-monomer input, with no failed outputs.

Each layer of the cascade contributes an additive share of the named output (Figure 6).

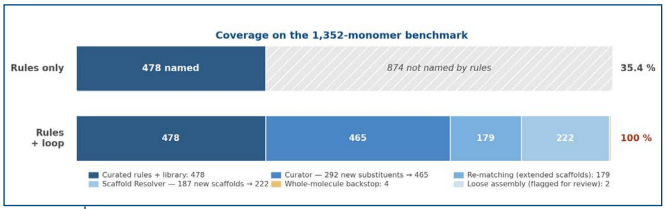


Figure 6: Coverage on the 1,352-monomer benchmark. The rule engine alone names 478 / 1,352 (35.4 %). The agentic loop layers — substituent-catalogue growth, new parent scaffolds, re-matching against extended scaffolds, and a small whole-molecule backstop — bring the total to 1,352 / 1,352 (100 %). Every valid single-monomer input is named and verified by structural read-back

HEPen — the whole-molecule backstop in action

A single benchmark input — 6-hydroxyethyl penicillanic acid, a β-lactam in the penicillin family — illustrates why the whole-molecule backstop is the mechanism that turns near-total coverage into a 100 % guarantee. The rule engine finds no clean parent scaffold and no catalogued substituent for the 1-hydroxyethyl side chain. The Scaffold Resolver's structural match flags the penicillanic-acid scaffold Pen at 81% stereo-blind MCS, but the stereo-aware structural check refuses to absorb it: HEPen's C6 stereocentre lives inside the β-lactam core ring rather than on a pendant group, and a core-level stereo difference cannot be expressed as a decoration on top of Pen. The extended scaffold library offers no better fit. The whole-molecule backstop then records the whole molecule as its own scaffold, by construction round-trips through the structural check, and the agent names it HEPen — HydroxyEthyl-Penicillanic — compact, HELM-style, and meaningful at a glance.

Catalogue growth and audit close the loop. Table 1 summarises each role's contribution to coverage and its model cost.

Role/Mechanism	Contribution	Model Cost
Curator	292 new substituents → 465 monomers	~1.0 M Sonnet Tokens
Reviewer	18 removed, 8 revised (quality, not coverage)	~0.3 M Sonnet Tokens
Scaffold Resolver	187 scaffolds → 403 monomers	~1.2 M Sonnet Tokens
Whole-molecule Backstop	4 monomers	—

Table 1: Per role / mechanism: what was added, what got named with it, and what it cost.

After the Reviewer's audit, **292 new substituents** are exportable and join the 210 curated baseline for a **502-entry shippable catalogue**. Every named output carries its origin (curated rules · run-discovered substituent · re-matched parent · new parent scaffold · loosely assembled · whole-molecule), and every catalogue entry carries its Reviewer verdict — a registration-grade audit trail.

SUMMARY

- The open sub-problem the paper frames — automatic naming of novel single-monomer residues that fall outside today's curated HELM catalogues — is closed. On 1,352 valid single-monomer residues mined from beyond the curated set, Helmsman names every input (1,352 of 1,352). Every name is verified by structural read-back.
- The mechanism is a rule engine plus an agentic loop — Curator, Reviewer, Scaffold Resolver — over the Hold (the running catalogue) and the same round-trip structural check, with a final **whole-molecule backstop** for residues irreducible to any existing scaffold. The freedom granted to the agents widens at each role (add substituents; remove them; re-match against a different parent scaffold and add new scaffolds; record the molecule itself as its own scaffold); the structural check bounds every step by requiring the emitted name to reconstruct the input.
- Helmsman complements rather than replaces the Pistoia ecosystem. Naming conventions are adopted from HELMify (Godfrey et al., 2026) without modification; new substituents and scaffolds enter the live catalogue only after passing the structural check and (for the exportable catalogue) the Reviewer.
- The output is registration-grade and auditable. Three name forms per residue; provenance per record; *needs_review* flags route to a triage queue for human catalogue review.

REFERENCES

1. Boccaletto, P., Stefaniak, F., Ray, A., Cappannini, A., Mukherjee, S., Purta, E., Kurkowska, M., Shirvanizadeh, N., Destefanis, E., Groza, P., Avşar, G., Romitelli, A., Pir, P., Dassi, E., Conticello, S. G., Aguilo, F., & Bujnicki, J. M. (2022). MODOMICS: A database of RNA modifications. 2022 update. *Nucleic Acids Research, 50*(D1), D231–D235. https://doi.org/10.1093/nar/gkab1083
2. Chen, W. L., Leland, B. A., Durant, J. L., Grier, D. L., Christie, B. D., Nourse, J. G., & Taylor, K. T. (2011). Self-contained sequence representation: Bridging the gap between bioinformatics and cheminformatics. *Journal of Chemical Information and Modeling, 51*(9), 2186–2208.
3. Cochrane, S. A., & Vederas, J. C. (2016). Enduracididine, a rare amino acid component of peptide antibiotics: Natural products and synthesis. *Beilstein Journal of Organic Chemistry*, *12*, 2156–2168.
4. Dunin-Horkawicz, S., Czerwoniec, A., Gajda, M. J., Feder, M., Grosjean, H., & Bujnicki, J. M. (2006). MODOMICS: A database of RNA modification pathways. *Nucleic Acids Research, 34*, D145–D149.
5. Gfeller, D., Michielin, O., & Zoete, V. (2013). SwissSidechain: A molecular and structural database of non-natural sidechains. *Nucleic Acids Research, 41*(D1), D327–D332. https://doi.org/10.1093/nar/gks991
6. Godfrey, A. G. et al. (2026). HELMify: Automated generation of HELM monomer identifiers from chemical structures. *ChemRxiv*.
7. He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., Sun, Y., Li, X., Dai, Q., Song, C.-X., Zhang, K., He, C., & Xu, G.-L. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science, 333*(6047), 1303–1307. https://doi.org/10.1126/science.1210944
8. IUPAC-IUB Commission on Biochemical Nomenclature. (1968). A one-letter notation for amino acid sequences. Tentative rules. *Journal of Biological Chemistry, 243*(13), 3557–3559.
9. Wang, Z. et al. (2016). Total synthesis of mannopeptimycins α and β. *Journal of the American Chemical Society, 138*. https://doi.org/10.1021/jacs.6b01384
10. Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., Veij, M. de, Ioannidis, H., Lopez, D. M., Mosquera, J. F., Magariños, M. P., Bosc, N., Arcila, R., Kizilören, T., Gaulton, A., Bento, A. P., Adasme, M. F., Monecke, P., Landrum, G. A., & Leach, A. R. (2024). The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research, 52*(D1), D1180–D1192. https://doi.org/10.1093/nar/gkad1004
11. Zhang, T., Li, H., Xi, H., Stanton, R. V., & Rotstein, S. H. (2012). HELM: A hierarchical notation language for complex biomolecule structure representation. *Journal of Chemical Information and Modeling, 52*(10), 2796–2806. https://doi.org/10.1021/ci3001925

APPENDIX — THREE-NAME SCHEMA REFERENCE

Form	Length	Composition
Traditional	Exactly 6 characters	Anchor first letter + 5-character deterministic hash
Long	≤ 16 Characters	Descriptive name with tiered compression; no hash suffix
Full	Unbounded	Raw pipeline output

Uniqueness across registry entries is preserved by the (traditional, long) tuple. Example outputs from the two case studies:

SMILES	Full	Traditional	Method
Nc1nc(=O)[nH]cc1C(=O)O	CYT5COOH	CPYSPK	Curator (substituent CO2H)
C1[C@H](NC(=N1)N)[C@@H](O)[C@@H](C(=O)O)N	R3OH4gCyc	RNLQM3	Scaffold Resolver (re-match + substituent gCyc)

Dassault Systèmes is a catalyst for human progress. Since 1981, the company has pioneered virtual worlds to improve real life for consumers, patients and citizens.

With Dassault Systèmes' **3DEXPERIENCE** platform, 370,000 customers of all sizes, in all industries, can collaborate, imagine and create sustainable innovations that drive meaningful impact.

For more information, visit: **www.3ds.com**

3DEXPERIENCE®

Europe/Middle East/Africa
Dassault Systèmes
10, rue Marcel Dassault
CS 40501
78946 Vélizy-Villacoublay Cedex
France

Asia-Pacific
Dassault Systèmes
17F, Foxconn Building,
No. 1366, Lujiazui Ring Road
Pilot Free Trade Zone, Shanghai 200120
China

Americas
Dassault Systèmes
175 Wyman Street
Waltham, Massachusetts
02451-1223
USA

Virtual Worlds for Real Life

DASSAULT SYSTEMES

©2026 Dassault Systèmes. All rights reserved. **3DEXPERIENCE**, the 3DS logo, the Compass icon, IFWE, 3DEXCITE, 3DVIA, BIOVIA, CATIA, CENTRIC PLM, DELMIA, ENOVIA, GEOVIA, MEDIDATA, NETVIBES, OUTSCALE, SIMULIA and SOLIDWORKS are commercial trademarks or registered trademarks of Dassault Systèmes, a European company (Societas Europaea) incorporated under French law, and registered with the Versailles trade and companies registry under number 322 306 440, or its subsidiaries in the United States and/or other countries. All other trademarks are owned by their respective owners. Use of any Dassault Systèmes or its subsidiaries' trademarks is subject to their express written approval.

DS-10082-0626